DOCUMENT RESUME

ED 373 631 HE 027 617

AUTHOR Jacobs, Stanley S.
TITLE Technical Characteristics and Some Correlates of the California Critical Thinking Skills Test, Forms A and B. AIR 1994 Annual Forum Paper.
PUB DATE May 94
NOTE 37p.; Paper presented at the Annual Forum of the Association for Institutional Research (34th, New Orleans, LA, May 29-June 1 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Freshmen; *Comparative Analysis; *Critical Thinking; Higher Education; Intelligence Tests; Predictive Measurement; *Test Construction; Test Format; Testing; Test Reliability
IDENTIFIERS *AIR Forum; *California Critical Thinking Skills Test (College); Test Equivalence

ABSTRACT

This study analyzed and tested the equivalence of Forms A and B of the California Critical Thinking Skills Test (CCTST). In designing the CCTST, Form A was composed of 34 items from a bank of 200. To develop a parallel measure, Form B was developed by rewriting 28 of the 34 items and rearranging their order. Study participants were all entering first-year students at a large private university. During other orientation activities 684 students completed Form A and 692 students completed Form B. Results of data comparison found that arithmetic means for the forms were significantly different indicating a lack of equivalence between forms. Principal component analysis and specific patterns of item intercorrelations differed between forms, with the lack of equivalence apparently due to the changes in Form A items which were carried out in order to create Form B items. Internal consistency reliabilities for total sub-test scores were uniformly low, and it appeared the CCTST scores largely reflect verbal intelligence of the type measured by the Student Aptitude Test (SAT). Analysis of the results concluded that CCTST may be acceptable for research purposes but not for decision-making concerning individual students, especially with respect to sub-test scores and score differences. (Contains 33 references and 11 tables of data). (Author/JB)

Technical Characteristics and Some Correlates of the California Critical Thinking Skills Test, Forms A and B

Stanley S. Jacobs

Office of Planning and Institutional Research

Villanova University

TECHNICAL CHARACTERISTICS AND SOME CORRELATES

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Fourth Annual Forum of the Association for Institutional Research held at The New Orleans Marriott, New Orleans, Louisiana, May 29, 1994 - June 1, 1994. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Editor
Forum Publications

Abstract

Forms A and B of the California Critical Thinking Skills Test (CCTST) were administered to two randomly-formed groups of undergraduate students at a large eastern university, as part of the freshman orientation process. Arithmetic means for the forms were significantly different indicating a lack of equivalence between forms. Principal component analyses and specific patterns of item intercorrelations differed between forms, with the lack of equivalence apparently due to the changes in Form A items which were carried out in order to create Form B items. Internal consistency reliabilities for total and sub-test scores were uniformly low, and it appears the CCTST scores largely reflect verbal intelligence of the type measured by the SAT. It was concluded that the CCTST may be acceptable for research purposes (e.g., as a blocking variable or covariate), but not for decision-making concerning individual students, especially with respect to sub-test scores and score differences.

Technical Characteristics and Some Correlates of the California Critical Thinking Skills Test

Prior to the appearance of the California Critical Thinking Skills Test (CCTST) (Facione, 1992), the two most commonly used standardized objectively-scored tests of critical thinking ability, appropriate for high school and college-age individuals, were the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980) and the Cornell Critical Thinking Test (Level Z) (Ennis, Millman & Tomko, 1985). Although the Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985) is contemporary with these instruments, the test requires the evaluation of constructed responses by raters who have completed at least a college-level course in informal logic, critical thinking, or the equivalent.

The Watson-Glaser was originally developed as a measure of the dependant variable in a major experiment designed to examine the effects of instructing high school students in critical thinking (Glaser, 1941). The Watson-Glaser is currently available in two equivalent forms, each containing five subtests, entitled Inference, Recognition of Assumptions, Deduction, Interpretation and Evaluation of Arguments. Reviews of the instrument, as well as extensive references, are provided by Berger (1985) and by Helmstadter (1985).

The Cornell was originally designed to assess critical thinking competencies in order to evaluate the effectiveness of curricular and instructional innovations, and (like the Watson-Glaser) is the result of a long-term program of research by the authors and others. The Cornell produces only a total score for the two levels of the test (Level X, generally intended for elementary and junior high students, and Level Z for high school and college students and adults). The total score is based upon items which measure induction, deduction, evaluation, observation, credibility of statements made by others, identification of

assumptions and the ability to discern meaning. Reviews and reference lists for the Cornell have been provided by Hughes (1992) and by Malcolm (1992).

Modjeski and Michael (1983) provided a comparative evaluation of these two instruments, based on the degree to which they met standards set forth in the 1974 Standards for Educational and Psychological Tests (American Psychological Association, 1974). Although the evaluations are based on forms of the two measures which have subsequently been revised (Helmstadter (1985, p. 1693), and Hughes (1992, p. 241), describe the revisions as slight), as were the Standards, neither the measures nor the Standards have changed in a fashion that would vitiate their conclusion that "...the Watson-Glaser was evaluated as a superior measuring device to the Cornell..." (Modjeski & Michael, 1983, p. 1196).

While Berger, Helmstadter, Hughes and Malcolm are unanimous in their cautions concerning the uses made of scores resulting from the Watson-Glaser and the Cornell, they all recommend the Watson-Glaser over the Cornell, primarily due to the availability of equivalent forms, data concerning technical characteristics and the extensive normative data for the Watson-Glaser.

The CCTST is an objectively-scored standardized instrument which addresses the cognitive skills dimension of critical thinking. The CCTST is the result of a conceptualization of critical thinking which emerged from a two year Delphi research project sponsored by the American Philosophical Association (Facione, 1990a). The panel of experts involved in the Delphi project included 46 persons active in critical thinking research, education and assessment. The panel's conceptualization of the critical thinking construct was summarized by Facione (1990a):

> We understand CT to be purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation and inference, as well as explanation of the evidential, conceptual, methodological, criteriological or contextual considerations upon which that judgement is based. (p.4).

The CCTST was subsequently constructed using a bank of 200 multiple-choice items. Through try-outs,

item analysis and revision, a set of 34 items was selected and designated Form A. The CCTST score is the simple sum of the number correct. However, the 34 items can be scored to yield three sub-scores, termed Analysis, Evaluation and Inference, or 30 of the items can be scored to yield two sub-test scores, termed Deductive Reasoning and Inductive Reasoning. Facione cites K-R_{20} (or alpha) coefficients for test scores based on all 34 items that range from .68 to .70 (Facione, 1992, p. 12), with the estimated reliability for the published Form A to be .70. No reliability estimates or intercorrelations are offered for the two suggested sets of sub-test scores.

Subsequently, Form B was developed by re-writing 28 of the 34 items appearing on Form A, by substituting different terms, names, concepts and contexts while attempting to maintain the type of topic or problem involved, and the specific CT behavior assessed by the original item. The original order of items on Form A, as well as position on page, item length and other "appearance factors" were maintained on Form B. Although the manual states 13 items are unchanged from Form A to Form B (p. 4), an examination of both forms reveals only six are unchanged: items number 13, 15, 31, 32, 33 and 34. However, the order of the alternatives for these items has been scrambled.

This procedure amounts to an attempt to construct a second, parallel (i.e. equivalent) measure on an item-by-item, deliberate basis, working from a fixed initial set of items dealing with specific content within a specific context, eliciting certain behaviors and producing certain statistical characteristics. A number of empirical studies, discussed by Cronbach (1984), have demonstrated the difficulties in developing parallel tests in a deliberate fashion, using expert judgement, as opposed to more conventional approaches (e.g., see Millman & Greene, 1993, pp. 348-349). Format effects have, on occasion, been shown to be particularly troublesome (e.g., see Green, 1984).

The equivalence of Forms A and B is argued on conceptual grounds, with within form data (e.g. measures of central tendency and variability) from two groups of undergraduates, ($N=90$), offered in support of the equivalence. The alpha reliability of Form B is estimated to be .71. Equivalence, however,

should be investigated with respect to at least four considerations (Cureton, 1958):

a) Are the resulting data equally variable?

b) Are the measures equally difficult?

c) Are the measures equally reliable?

d) Are the measures assessing the same function or combination of functions (e.g. Are they factorially similar?)

These four considerations are basic to the classic view of "parallel forms". There are several alternative conceptualizations of the idea of parallel measurement procedures which do not require that the above considerations be wholly satisfied (Feldt and Brennan, 1993). However, the equivalence of CCTST forms A and B is argued from what appears to be a strictly classical perspective (Facione, 1992, p. 11), and the evaluation of their equivalence will involve Cureton's four considerations. The implications of alternative models for parallel forms will also be considered, however.

With reference to the question of validity, the perspective adopted will be that of Messick's "unified view" of test validity (Messick, 1980, 1981, 1988, 1993) The concept of validity has evolved from a focus on studies concerned with "types of validity", usually content validity, criterion-related validity and construct validity (American Psychological Association, American Educational Research Association and National Council on Measurement in Education, 1985), to a recognition that it was the "inferences drawn from test scores that were on trial" (Angoff, 1988, p. 30). Rather than an emphasis on "critical studies", or single statistical tests or indices, the current conceptualization of validation as a process includes "...the simultaneous consideration of the theoretical and empirical rationale of the test, the intended and actual uses of the test, the evidence verifying the construction of the test and the ethical implications of using the test" (Pittenger, 1993, p. 467).

An important consequence of the broadened perspective concerning validity is that the validation process is conditional. One's conclusions are dependent upon the purposes for which test data are obtained. For

example, a test may possess certain psychometric qualities that would allow the use of test scores as a blocking variable or a covariate in an experimental design. These same qualities may lead a test user to conclude that the test data should not be used to evaluate individuals, or to make comparisons among individuals. Another consequence of a more unified view of the concept of validity is that the arbitrary distinction between the concepts of reliability and validity is becoming blurred. A more inclusive concept, akin to the idea of utility, seems to be emerging.

The investigation of validity, then, must be regarded as an ogoing process with each study producing limited information concerning specific uses or interpretations of the data resulting from a measurement procedure. The present study is intended to examine some of the technical characteristics and correlates of the CCTST, to begin to accumulate the information needed to answer the question of the utility of CCTST data.

Method

Subjects

Participants in the study were all entering first-year students (N=1383) at a large Eastern private university who appeared for regularly-scheduled freshman orientation activities during August 1993. The CCTST was completed during a two-day orientation session in conjunction with a number of other experiences. The average age was 18.7 years. The average percentile rank for SAT Verbal scores was 51.3 and for SAT Quantitative scores was 58.6. The sample was 52 percent male.

Procedure

The CCTST data were collected from all students as part of the administration of a battery of tests and other data collection procedures. Students were told that the data were collected to enable the University to incorporate the information into research intended to improve University programs. Confidentiality of data was emphasized.

As part of the University's freshman orientation procedure, 75 orientation counselors were trained to coordinate the various activities scheduled. The CCTST was scheduled to be administered (with other measures) during this period to small groups (not larger than 25) led by a trained counselor. Each counselor was familiarized with the basic procedures of standardized test administration and was provided a set of Test Administration Instructions. Students were randomly assigned to groups and test administration occasions. Orientation counselors administered randomly selected packets containing Form A or Form B to their group. This procedure resulted in 684 students completing Form A and 692 completing Form B; seven students were deleted from the analysis due to incorrect or missing student identification numbers.

The CCTST was administered according to procedures recommended in the manual. Scannable answer sheets were employed (Computest number 16412), and all analyses were carried out manually or with SPSS procedures.

Results

Descriptive statistics concerning Form A and Form B data are presented in Table 1.

Insert Table 1 About Here

As a first test of the equivalence of Forms A and B, the variances of the scores produced by Forms A and B were contrasted using Cochran's C statistic (Winer, 1962). The resulting statistic of .518 has a chance probability of .34; the difference in variances is not significant.

A one-way analysis of variance (ANOVA) was then carried out, contrasting the 684 students who had completed Form A with the 692 students who had completed Form B, but with respect to their scores on only the six items unchanged and common to both A and B. The resulting F (1, 1374) of 2.68 is not significant ($p > .05$). In other words, the level of performance on Forms A and B was equivalent with respect to the six common items. The lack of a significant difference in level of performance on the six items common to Forms A and B, and the close correspondence between those students who completed Form A and those who completed Form B, with respect to the descriptive variables of SAT scores, age and gender proportions, provides a perspective for analyzing the difference in performance over the 28 items which were modified in the creation of Form B.

Mean scores based on the 28 modified items on Form A and Form B were analyzed with a one-way analysis of covariance, with scores based on the six common items as the covariate. The resulting F (1,373) of 6.30 is significant. ($p < .01$). This indicates that the mean for Form A, based on the 28 modified items, is significantly higher than the mean for Form B.

To summarize these analyses, it appears that the modifications made on Form A items to transform them into Form B items have made the items more difficult, resulting in two test forms which are not equivalent with respect to difficulty. The nature of the effect was further investigated by examining the

Pearson product-moment correlation between item difficulties for Form A and for Form B, for the six items unchanged and common to both forms, and for the 28 items which were modified for use on Form B. Two possibilities exist: Form A may be systematically easier on an item-by-item basis, or the effect of the modifications may vary from item to item, producing a more difficult Form B overall, with the effect on item difficulty varying with specific items.

The correlation of Form A and Form B item difficulties for the six items unchanged and common to both forms was found to be .98, an almost perfect correspondence. The correlation between item difficulties for the 28 modified items on Form A and Form B was found to be .83. Although both coefficients indicate statistically significant ($p < .01$) and strong relationships between forms, a z-test on the difference between the correlations indicates a significant difference ($z = 1.88$; $p < .05$). This indicates that not only is Form A significantly easier than Form B (as a result of the 28 modified items), but there is a significant difference between forms in the relative positions of the modified items with respect to the continuum of item difficulty.

The various proposed sub-tests were next examined. Summary statistics for the five proposed subtests for Form A and Form B are presented in Table 2.

Table 2 About Here

The entry in the table labeled Alpha (34) is simply the calculated alpha reliability for a subtest "stepped-up" using the Spearman-Brown formula to estimate the alpha reliability if the subtest were composed of 34 items. Since the subtests vary in terms of number of items, this standardizes the alphas with respect to subtest length, and allows comparisons of alphas among subtests and with total test alphas.

The Pearson product-moment intercorrelations among the five possible subtests for Forms A and B are

presented in Table 3. The upper half of the correlation matrix (above the diagonal) represents relationships among Form A subtests; the lower half represents Form B relationships. Since the two sets of subtests on each form are simply re-configurations of the same set of item scores, the statistics for the two sets of subtests are not independent. For example, 10 of the 14 items comprising the Evaluation subtest would be included in the Induction subtest, which contains only 14 items total. One would expect these two subtests to have similar characteristics, as well as a substantial intercorrelation.

Insert Table 3 About Here

To examine the correspondence between the pattern of relationships among subtests for Forms A and B, the rank order correlation between the subtest-pair correlations for Forms A and B was computed. The correlation is .99, almost perfect. Although specific numeric values of the strength of relationship vary somewhat, when specific pairs are considered, there is very close correspondence in the pattern of relationships among subtests. It can be noted that the average intercorrelation among subtests is somewhat higher for Form B than Form A (.48 versus .42). This can be predicted from the higher alpha reliability for Form B. Alpha is a direct reflection of the average intercorrelation among the total set of items comprising a measure. If alpha is higher, than the average intercorrelation is higher, which would tend to make the average intercorrelation among partitionings of those items higher. Although the total test score alpha coefficient for Form B is numerically higher than for Form A (see Table 1), the difference is not statistically significant. Feldt's W (Feldt, 1969) of 1.07, with 683 and 691 df, is not significant ($p > .05$).

A consideration of the subtest reliability estimates in Table 2, and the pattern of intercorrelations in Table 3 indicates that some sets of items are measuring some aspects of behavior in a reliable manner (e.g.

Deduction) while others are not (e.g. Analysis). To attempt to determine the basis of the differences in alpha estimates among subtests, student performance on an item-by-item basis was examined, first considering the full set of 34 items and then by examining the relationships among items making up the various subtests.

The CCTST is composed of 19 four-option multiple choice items and 15 five-option multiple choice items. Chance level performance can be estimated by considering the probability of guessing at the answer to an objectively-scored item and answering it correctly. Four option items give a student a one in four opportunity to guess the correct answer; five options involve a one in five opportunity. Considering the number of each type of item (or opportunity) indicates chance level scores on the CCTST would be equal to seven or eight. In fact, higher (and lower) scores are possible. Chance level scores may result from less-able students attempting very difficult items; they simply cannot answer very many correctly. They may also result from essentially random behavior. The problem lies in attempting to distinguish between honest attempts and random behavior. Facione, incidentally, gives a CCTST score of eight a percentile rank of 4, indicating 4 percent of the norming sample scored at or below an expected chance score value.

Item means (which for dichotomously-scored items is simply the proportion answering the item correctly) were examined for Forms A and B. Using as the criterion for chance-level performance the expected proportion of students getting an item correct through blind guessing (.25 for 4 option items; .20 for 5 option items), five items (numbers 1, 9, 21, 22, and 25) were found to have means at or below the chance level on Form A. These same items, as modified, were found to have below-chance means on Form B (See Table 4).

Insert Table 4 About Here

In an attempt to determine the basis for the chance-level item means (guessing or only a few of the most-able students answering the item correctly) the item-total test score correlations were examined. Since logic would demand that those students with the higher levels of CT ability would earn higher scores on the CCTST, there will be a positive relationship between item and test performance if the item is so difficult only the better students answer it correctly.If there is no relationship, then one could conclude that students are guessing at the item, misconstruing the item, the options are ambiguous or the like. In any event, an item which is independent of the total test score simply makes the total test score more difficult to interpret. Negative correlations indicate that failing an item is associated with doing better over the total test (and vice-versa). Adding items like these to a set of positively-correlated items both depresses internal consistency reliability estimates and produces test scores which are difficult to interpret. A reliable negative relationship between an item and total test score may indicate a systematic problem with the item (e.g., disagreement between test author and respondents concerning the keyed-correct response).

Corrected item-total test score correlations for Forms A and B are presented in Table 5.

Insert Table 5 About Here

Items 9 and 21 (of the five items with low item means) show low correlations with total test score. Coupled with the low means for these two items, one might conclude students are, in fact, guessing at these items. The other three items (numbers 1, 22 and 25) appear to be more discriminating, although number 25 appears marginal. More bothersome are items 11 and 13, which show negative correlations, and a number of others which show very low relationships. Considered individually, one can examine the effect on the alpha reliability as these items are removed from the set of 34. Items with low or negative correlations with total test score have an adverse effect on alpha; their deletion has a positive

effect on alpha for the remaining set of 33. One might add that their deletion may have a positive effect on the interpretability of the CCTST total test score.

Many of the same problems are shared between forms; items 9 and 21 have low item-test correlations on both forms. Item 11 is consistently negatively correlated; Item 33 appears to be independent of total test score. Overall, the pattern of inconsistent item-total test score correlations calls into question the interpretation of a test score based on the simple sum of those item scores.

If one shifts from the concept of a meaningful total test score to the concept of meaningful subtest scores, each of which, it could be argued, should be reliable (therefore potentially meaningful) and relatively independent of other subtests, then the pattern of varied item-total test score correlations could be considered irrelevant. One could argue that the pattern of varied relationships will be clarified if the items are sorted into appropriate categories (i.e. Induction, Deduction; Analysis, Evaluation, Inference). The patterns of item-subtest score intercorrelations for these classifications of items are presented in Tables 6 and 7, as are the average item-intercorrelations among the sets of items making up the several suggested subtests.

Insert Table 6 About Here

Insert Table 7 About Here

The pattern of subtest reliabilities is summarized in Table 2. While some item classifications produce scores of such low reliability as to be virtually worthless (e.g., Analysis), a number of others produce

data which is sufficiently reliable for research purposes (e.g., Deduction, Inference). The reliability of some item classifications is higher than the total test reliability, when corrected for the effect of the differing numbers of items, supporting some of the item classifications. However, the large number of correlation coefficients in Tables 6 and 7 that are less than .100 indicate that each subtest is measuring more than one behavior. These behaviors should be regarded as largely independent of whatever the strongly inter-related items are measuring. The negative correlations are (for the most part) weak, and probably reflect independent behaviors or skills, or (simply) unreliable measurements. The low (less than .10) average item intercorrelations result in depressed alpha reliabilities. As Cronbach (1951) pointed out, alpha is actually the average item-intercorrelation "stepped-up" by the Spearman-Brown formula to estimate the reliability of the full-length test composed of those items.

To put the issue in more pragmatic terms, subtest scores are interpretable only to the degree to which items which are summed to produce the score "hang together". In the event that a subtest score is based on items which are largely independent of one another (or negatively related to each other), the subtest score reflects several behaviors with the relative strength of each unknown. In short, there is a problem in interpreting what the subtest score means, and what differences between or among subtests (i.e., profiles for students) suggest.

To further investigate the legitimacy of the sorting of CCTST items into subtests (either the Induction-Deduction classification or the Analysis-Evaluation-Inference classification) an exploratory principal components analysis was carried out on the item data for Form A and Form B. Both analyses produced 14 clusters of linearly-combined items with eigenvalues greater than 1.00. This indicates that there are 14 clusters of intercorrelated items, where the clusters are independent of each other, with the 14 independent sets of items accounting for just over 50 percent of the variance in CCTST scores. In short, it could be argued that all CCTST scores reflect the influence of a number of behaviors, that the CCTST total score is based upon a large number of heterogeneous behaviors, and that the classification of items

into subtests is not supported by the manner in which students respond to those items.

Using the arbitrary criterion of a loading of .30 or greater, the specific items associated with each component, the items' loadings, the eigenvalues for each component and the cumulative percentage of variance accounted for are summarized in Table 8 for Form A and Table 9 for Form B.

Insert Table 8 About Here

Insert Table 9 About Here

It is apparent that neither principal components analysis supports the recommended classifications of CCTST items. Using the same procedure and criteria for the evaluation of item interrelationships, two dissimilar groups of items seem to emerge for the two test forms analyzed.

In an attempt to identify the reason for the difference in grouping of items, the correlation matrices summarizing the intercorrelations among items on Forms A and B were examined. It was found that the pattern of significant item intercorrelations ($p \leq .01$) was dramatically different for each of the forms, producing the differences in the analyses.

For example, Item 1 on Form A is correlated significantly with Items 2, 3, 4, 6, 10, 14, 22, 23, and 27. On Form B, the same item is correlated significantly with items 3, 8, 14 and 26; only two items from both Forms A and B (numbers 3 and 14) are consistently related to Item 1.

Four additional items were selected at random from the remaining 27 items which differ between Forms A and B. The significant relationships of these items with the remaining items were identified through an inspection of the item intercorrelation matrices for Forms A and B; the significant relationships within forms are summarized in Table 10. The items selected at random were numbers 6, 10, 14 and 24.

Insert Table 10 About Here

Bearing in mind that Form A and Form B items are intended to be equivalent, producing alternate forms of the CCTST, through the deliberate modification of what were probably regarded as inconsequential features of Form A items, the lack of similarity in the pattern of relationships is striking. Item 6 is significantly correlated with 14 other items. However, only nine of the 14 are common to both Forms A and B. Item Number 10 is significantly correlated with 12 other items; only two are common to both forms. Item Number 14 is correlated significantly with 16 other items; 11 are common to Forms A and B. Item number 24 is correlated significantly with 15 other items, 12 of which are the Form B version only. Only one item out of 15 shows a similar relationship with item Number 24 on both forms of the CCTST.

The pattern of item intercorrelations among the six items which were unchanged between forms was examined next. It was found that Item 13 was independent of the remaining five items on both Forms A and B. Item 15 was significantly correlated with Item 34 on Form A only, and independent of the remaining four items on both forms. Item 31 is significantly correlated with Items 32, 33 and 34 on both forms. Item 32 is correlated with Items 31 and 34. Item 33 is correlated with items 31 and 34 on Form A but only Item 31 on Form B. Item 34 is correlated with Items 15, 31, 32, and 33 on Form A and Items 31 and 32 on Form B. It appears the correspondence in item performance is much closer for the unchanged than for the modified items, although not perfect.

The pattern of item intercorrelations supports the principal component results; it is probably safe to say that the pattern of item interrelationships between modified items is highly idiosyncratic to test Form. Since the sample of items was drawn in a random fashion, the same pattern is probably similar for the

23 remaining modified items not examined.

To investigate the effect of selected demographic variables on CCTST performance, a stepwise regression analysis was carried out for those students who had completed Form A and for those who had completed Form B. Scholastic Aptitude Test (SAT) scores (verbal and quantitative), age and gender were entered as predictors and the six scores obtainable from the CCTST (Total, Deduction, Induction, Analysis, Evaluation and Inference) were regarded as criteria in six successive stepwise regression analyses. In every analysis, SAT verbal scores were the best predictor. With the variance in CCTST performance attributable to SAT verbal scores statistically removed, the second best predictor in every instance except two was the SAT quantitative score. In the prediction of Analysis scores, age was a better predictor than SAT quantitative scores but only for Form B. SAT verbal scores were the only predictor significantly related to Analysis scores on Form A, which might be expected considering the very low reliability of Analysis scores. In the prediction of Form A Induction scores, gender was a better second predictor than SAT quantitative scores. The pattern of simple and multiple correlations for the best and second-best predictors of the various criteria for Form A and Form B is summarized in Table 11.

Table 11 About Here

It must be remembered that the regression analyses summarized in Table 11 are not independent within forms. Subtest scores of the two types are simply recombinations of the same items and the total score is composed of all 34 items. However, considering the alpha reliabilities of Forms A and B, and the correlations between the several divisions of items and the verbal score from the SAT, one is led to the conclusion that whatever the CCTST is measuring is largely whatever the SAT verbal score represents.

Discussion

Perhaps the most important finding of the present study is not the lack of equivalence of Forms A and

B with respect to difficulty. As Feldt and Brennan (1993) point out, essentially tau-equivalent forms may exhibit differences in arithmetic means and in variances. However, the difference in means between forms should be reflected in a similar difference between individuals' true scores on the two forms. A person's true score on Form A would be related to that person's true score on Form B through some non-zero constant value, C_{AB}. If this were the case, an equating process could be implemented to statistically equate data from Forms A and B (e.g., see Petersen, Kolen and Hoover, 1993). Unfortunately, the differences between forms seem to be much more complex. The differences in relative item difficulty within forms is an indication that the effects of item modification are complex, and vary from item to item. The fact that the principal components analysis produced two dissimilar groups of items, neither of which correspond to suggested categorizations of items is perplexing. The finding that item interrelationships are extremely inconsistent between forms indicates that item analysis and revision strategies used with one form will probably not have the same effect with the same item(s) on the second form. This finding effectively rules out the most liberal approach to estimating the reliability of the CCTST, the model of multi-factor congeneric forms (Feldt & Brennan, 1993, p. 111).

Taken together, the findings that the forms differ with respect to difficulty, that the Form B item modifications have effected the relative difficulty of items, that the tests appear to be factorially complex and the complexity differs between forms, and that the pattern of item relationships is largely idiosyncratic to form, all make CCTST score interpretation highly tentative. Neither total test score nor subtest scores should be regarded as measuring unitary constructs. Total test scores may be found to be adequate for research purposes, but their interpretability prevents their use in making decisions concerning individuals. Regardless of what each possible CCTST score might be measuring, the comparability of scores from Forms A and B is highly questionable.

The analyses of Form A and Form B subtests indicates that some subtests are somewhat more homogeneous, therefore interpretable, than the total test score, (e.g., Deduction), while others are not

(e.g., Analysis). One would hope for subtests composed of moderately intercorrelated items with the subtests at least relatively independent of each other. This would allow one to speak of meaningful subtest scores, and of differences between subtests. Unfortunately this does not appear to be the case. With the test reliabilities "stepped up" to estimate a reliability based on 34 items, some subtests appear to be at least as reliable as total test scores. This, however, does not reflect the actual nature of the data, which are based on smaller numbers of items. The failure of the principal component analysis to support the proposed CCTST item classifications (and the apparent between-form differences) renders their use highly questionable. Their low reliabilities and relatively high correlations (given the level of reliability involved) would prevent their use in making decisions concerning individuals, especially where one might be concerned with differences between or among subtests.

The high relationship between all suggested CCTST subtest scores, as well as total test scores, and SAT verbal scores also raises serious questions about the nature and extent of the unique contribution CCTST scores may make to our understanding of student behavior. This problem is not unique to the CCTST, however, and has been noted as a problem with other instruments of this type (e.g., Hughes, 1992, pp. 242-243). Additional studies need to be done to determine the degree to which CCTST scores represent relatively stable constructs (e.g., verbal intelligence, reading comprehension, affective variables influencing decision-making behavior), and to what extent CCTST scores may detect changes brought about by instruction or other experiences designed to modify critical thinking skills.

Facione has presented extensive data concerning the CCTST, as a supplement to the Manual (Facione, 1990b, 1990c, 1990d, 1990e). However, many, if not most, of the studies offered in support of the instrument are quasi-experimental in nature (Cook & Campbell, 1979, pp. 95-146), which produced data open to alternative interpretations. Substantial gaps exist in the normative and technical data for the CCTST.

The results of this study illustrate the difficulties in developing measures of complex behaviors. The

differences between forms are apparently largely due to what one might describe as "context effects" or "format effects". Apparently minor and harmless changes in item wording, phrases, nouns and so forth had a significant and complex effect on students' responses to the modified items. The resulting measures reveal how little is known about how much of a difference in a stimulus (i.e. the item) is required to produce a significant effect on the response.

References

American Psychological Association. (1974). Standards for educational and psychological tests. Washington, DC: Author.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Angoff, W.H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun (Eds.) Test validity (pp. 19-32). Hillsdale, NJ: LEA.

Berger, A. (1985). Review of Watson-Glaser Critical Thinking Appraisal. In J.V. Mitchell (Ed.), Ninth mental measurements yearbook (pp. 1692-1693). Lincoln, Nebraska: University of Nebraska, Lincoln, Buros Institute of Mental Measurements.

Cook, T.D. & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton-Mifflin Company.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L.J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

Cureton, E.E. (1958). The definition and estimation of test reliability. Educational and Psychological Measurement, 18, 715-738.

Ennis, R.H., Millman, J. & Tomko, T.N. (1985). Manual, Cornell critical thinking essay test. Pacific Grove, CA: Midwest Publications.

Ennis, R.H. & Weir, E. (1985). Manual, the Ennis-Weir critical thinking essay test. Pacific Grove, CA: Midwest Publications.

Facione, P.A. (1990a). Executive summary: Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Millbrae, CA: California Academic Press.

Facione, P.A. (1990b). The California Critical Thinking Skills Test - College Level; Experimental validation and content validity (Report No. 1). Millbrae, CA: California Academic Press.

Facione, P.A. (1990c). The California Critical Thinking Skills Test - College Level; Factors predictive of ct skills (Report No. 2). Millbrae, CA: California Academic Press.

Facione, P.A. (1990d). The California Critical Thinking Skills Test - College Level; Gender, ethnicity, major, ct, self-esteem, and the CCTST (Report No. 3). Millbrae, CA: California Academic Press.

Facione, P.A. (1990e). The California Critical Thinking Skills Test - College Level; Interpreting the CCTST, group norms, and sub-scores. (Report No. 4). Millbrae, CA: California Academic Press.

Facione, P.A. (1992). Manual, the California critical thinking skills test, form A and form B. Millbrae, CA: California Academic Press.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373.

Feldt, L.S. & Brennan, R.L. (1993). Reliability. In R.L. Linn (Ed.) Educational measurement (3rd ed.) (pp. 105-146). Phoenix, AZ: Oryx Press.

Glaser, E.M. (1941). An experiment in the development of critical thinking. New York: Teachers College of Columbia University Bureau of Publications.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. Educational and Psychological Measurement, 44, 551-561.

Helmstadter, G.C. (1985). Review of Watson-Glaser Critical Thinking Appraisal. In J.V. Mitchell (Ed.), Ninth mental measurements yearbook (pp. 1693-1694). Lincoln, Nebraska: University of Nebraska, Lincoln, Buros Institute of Mental Measurements.

Hughes, J.N. (1992). Review of Cornell Critical Thinking Tests. In J.J. Kramer & J.C. Conoley (Eds.) Eleventh mental measurements yearbook. (pp. 241-243). Lincoln, Nebraska: University of Nebraska, Lincoln, Buros Institute of Mental Measurements.

Malcoln, K.K. (1992). Review of Cornell Critical Thinking Tests. In J.J. Kramer & J.C. Conoley (Eds.) Eleventh mental measurements yearbook. (pp. 243-244). Lincoln, Nebraska: University of Nebraska, Lincoln, Buros Institute of Mental Measurements.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braum (Eds.), Test validity (pp 35-45). Hillsdale, NJ: LEA.

Messick, S. (1993). Validity. In R.L. Lynn (Ed.), Educational measurement (3rd ed.), (pp. 13-103). Phoenix, AZ: Oryx Press.

Millman, J. & Greene, J. (1993). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.) Educational measurement. (3rd ed.) (pp. 335-336). Phoenix, AZ: Oryx Press.

Modjeski, R.B. & Michael, W.B. (1983). An evaluation by a panel of psychologists of the reliability and validity of two tests of CT. Educational and Psychological Measurement, 43, 1187-1197.

Petersen, N.S., Kolen, M.J. & Hoover, H.D. (1993). Scaling, norming and equating. In R.L. Linn (Ed). Educational measurement. (3rd ed.). (pp. 221-262). Phoenix, AZ: Oryx Press.

Pittenger, D.J. (1993) The utility of the Myers-Briggs Type Indicator. Review of Educational Research, 63, 467-488.

Watson, G. & Glaser, E.M. (1980). Manual, Watson-Glaser critical thinking appraisal, forms A and B. San Antonio: Psychological Corporation.

Winer, B.J. (1962). Statistical principles in experimental design. New York: McGraw-Hill Book Co.

Table 1

Descriptive Data Concerning Research Participants Completing Form A and Form B, CCTST

Index	Form A	Form B
Number of participants	684	692
Mean Age	18.8	18.6
Mean SAT Verbal percentile	51.2	51.4
Mean SAT Quantitative percentile	58.5	58.6
Gender distribution	52.0 % male	52.0 % male
Mean, total test score	16.01	15.36
Mean, 28 altered items	13.03	12.50
Mean, 6 unaltered items	2.98	2.87
Median, total test score	16.00	15.00
Mode, total test score	16.00	14.00
Range of total test scores	5 - 27	5 - 29
Standard deviation, total test score	3.94	4.09
Alpha, total test score	.56	.59

Table 2

Summary Statistics for Form A and Form B Subtests, CCTST

Index	Induction (k=14)	Deduction (k=16)	Analysis (k=9)	Evaluation (k=14)	Inference (k=11)
	Form A Subtests				
Arithmetic Mean	6.48	7.66	4.54	5.78	5.70
Range of Scores	0-12	2-14	1-9	0-14	1-10
Standard Deviation	2.22	2.46	1.39	2.31	1.79
Median	7.00	8.00	5.00	6.00	6.00
Mode	7.00	7.00	5.00	6.00	6.00
Alpha	.42	.50	.04	.45	.36
Alpha (34)	.64	.68	.14	.67	.64
	Form B Subtests				
Arithmetic Mean	6.16	7.47	4.31	5.78	5.27
Range of Scores	1-11	1-14	1-9	0-11	0-10
Standard Deviation	2.11	2.51	1.48	2.10	1.88
Median	6.00	7.00	4.00	6.00	5.00
Mode	6.00	7.00	4.00	5.00	5.00
Alpha	.35	.53	.16	.33	.42
Alpha (34)	.56	.71	.42	.54	.70

Table 3

Relationships Among Form A and Form B Subtests, CCTST

Subtest	Induction	Deduction	Analysis	Evaluation	Inference
Induction	----	.23**	.22**	.83**	.38**
Deduction	.34**	----	.43**	.49**	.77**
Analysis	.31**	.49**	----	.20**	.20**
Evaluation	.83**	.54**	.26**	----	.34**
Inference	.49**	.80**	.30**	.42**	----

Note: Entries above the diagonal represent data from Form A. Those below the diagonal refer to Form B.

** $p < .001$.

Table 4

Item Means for Five Items Indicating Chance-Level Performance

Item No.	No. Options	Form A Item Mean	Form B Item Mean
1	4	.221	.150
9	5	.186	.152
21	5	.105	.140
22	5	.197	.153
25	4	.231	.210

Table 5

Item and Item-test Statistics for Forms A and B, CCTST

Question Number	Item Mean		Corrected Item-Total r		Alpha, Item Deleted	
	Form A	Form B	Form A	Form B	Form A	Form B
1	.221	.150	.184	.164	.548	.585
2	.325	.431	.143	.078	.552	.593
3	.401	.611	.154	.192	.550	.581
4	.287	.400	.189	.162	.547	.584
5	.716	.727	.054	.080	.562	.592
6	.718	.591	.227	.254	.542	.574
7	.759	.292	.096	.070	.557	.593
8	.680	.762	.080	.243	.559	.577
9	.186	.152	.025	.029	.563	.595
10	.330	.399	.159	.186	.550	.582
11	.282	.396	-.086	-.075	.576	.609
12	.504	.649	.164	.218	.549	.578
13	.360	.344	-.032	.068	.572	.594
14	.664	.571	.313	.309	.532	.568
15	.547	.515	.147	.175	.551	.583
16	.617	.655	.009	.185	.567	.582
17	.923	.910	.186	.169	.550	.585
18	.529	.555	.141	.234	.552	.576
19	.301	.285	.103	.117	.556	.589
20	.700	.542	.141	.235	.552	.576
21	.105	.140	.076	.051	.558	.593
22	.197	.153	.253	.238	.541	.579
23	.380	.259	.261	.227	.538	.578
24	.738	.686	.147	.218	.551	.578
25	.231	.210	.141	.059	.552	.594
26	.597	.515	.254	.257	.538	.574
27	.366	.353	.246	.330	.539	.566
28	.307	.251	.123	.108	.554	.589
29	.604	.559	.258	.251	.538	.574
30	.364	.295	.084	.065	.559	.594
31	.380	.311	.142	.163	.552	.584
32	.578	.594	.156	.102	.550	.591
33	.417	.418	.056	-.002	.562	.602
34	.702	.686	.240	.116	.541	.589

Table 6

Summary of Item-Subtest Score Correlations and Average Item Intercorrelations for the Induction-Deduction Classification of CCTST Items - Form A and Form B

Item No.	Induction Subtest		Item No.	Deduction Subtest	
	Form A	Form B		Form A	Form B
3	.100	.124	1	.190	.177
13	-.017	.052	2	.206	.118
20	.089	.209	4	.233	.159
21	.068	.075	5	.076	.077
24	.152	.149	6	.230	.281
25	.080	.028	8	.067	.223
26	.246	.192	9	.057	.047
28	.140	.003	14	.309	.311
29	.300	.228	15	.156	.179
30	.039	.034	16	.031	.173
31	.168	.190	17	.164	.187
32	.250	.116	18	.171	.221
33	.101	.040	19	.087	.109
34	.255	.073	22	.279	.225
			23	.226	.229
			27	.225	.257
Average Inter-Item r	.05	.03		.06	.07

Table 7

Item-Subtest Score Correlations and Average Item Intercorrelations for the Analysis-Evaluation-Inference Classification of CCTST Items - Form A and Form B

Item No.	Analysis Subtest Form A	Analysis Subtest Form B	Item No.	Evaluation Subtest Form A	Evaluation Subtest Form B	Item No.	Inference Subtest Form A	Inference Subtest Form B
5	.022	.031	1	.163	.114	14	.242	.271
6	.047	.108	2	.067	.035	15	.143	.131
7	.013	.021	3	.169	.169	16	.034	.158
8	-.023	.118	4	.140	.090	17	.157	.219
9	-.012	.012	25	.098	.010	18	.134	.209
10	.078	.092	26	.239	.166	19	.117	.088
11	-.046	-.031	27	.172	.230	20	.099	.182
12	.068	.123	28	.126	.032	21	.022	-.028
13	-.033	.017	29	.250	.208	22	.203	.171
			30	.077	.062	23	.162	.199
			31	.176	.159	24	.136	.153
			32	.160	.059			
			33	.100	.030			
			34	.225	.081			
Average Inter-item r	.004	.020		.055	.033		.050	.062

Table 8

Components, Eigenvalues and Cumulative Percentage of Variance Accounted for in Principal Components Analysis of Form A, CCTST

Component	Form A Items and Loadings	Eigenvalue	Cumulative % Variance Accounted For
1	6(.40); 14(.50); 22(.40); 23(.42); 1(.33); 17(.29); 26(.41); 27(.42); 29(.38); 34(.37); 4 (.32)	2.60	7.7
2	29(.35); 32(.57); 34(.41); 24(.34); 26(.30)	1.59	12.3
3	15(.35); 16(.45); 2(.31)	1.34	16.3
4	7(.37); 15(.40); 18(.37)	1.31	20.1
5	5(.40); 13(.45); 10(.33); 28(.31); 33(.33)	1.27	23.9
6	11(.35); 25(.35); 20(.30)	1.24	27.5
7	3(.45); 21(.46); 4(.32)	1.19	31.0
8	9(.38); 19(.39)	1.16	34.4
9	21(.40)	1.13	37.8
10	13(.33); 19(.33)	1.10	41.0
11	28(.38)	1.08	44.2
12	7(.34)	1.07	47.3
13	11(.40)	1.04	50.4
14	1(.33); 8(.33); 33(.34); 11(.30)	1.00	53.3

Table 9

Components, Eigenvalues and Cumulative Percentage of Variance Accounted for in Principal Components Analysis of Form B, CCTST

Component	Form B Items and Loadings	Eigenvalue	Cumulative % Variance Accounted For
1	4(.50); 6(.57); 8(.46); 14(.46); 15(3.5); 26(.49)	2.78	8.2
2	10(.55); 20(.37); 22(.38); 27(.64); 28(.30)	1.44	12.4
3	17(.69); 18(.70)	1.41	16.6
4	24(.31); 29(.42); 32(.68)	1.38	20.6
5	21(.45)	1.29	24.4
6	8(.35); 20(.30); 38(.38); 31(.78)	1.24	28.1
7	11(.33); 12(.54); 24(.32); 25(.70)	1.20	31.6
8	13(.65); 16(.58)	1.15	35.0
9	23(.43); 33(.78)	1.13	38.3
10	9(.62); 15(.38); 2(.63)	1.10	41.6
11	30(.76)	1.08	44.7
12	19(.78)	1.05	47.8
13	5(.72); 22(.32)	1.03	50.9
14	7(.77)	1.01	53.8

Table 10

Pattern of Significant Correlations ($p \leq .01$) Between Four Randomly Selected Items and Remaining CCTST Items, Within Forms A and B

Item pairs	Item No. 6		Item pairs	Item No. 10	
	Form A	Form B		Form A	Form B
6 & 1	.091	.136	10 & 1	.091	NS
6 & 4	.117	.146	10 & 3	NS	.105
6 & 8	NS	.183	10 & 11	-.109	NS
6 & 14	.159	.152	10 & 12	.106	.111
6 & 15	.095	.115	10 & 14	.099	.098
6 & 17	NS	.110	10 & 18	NS	.124
6 & 18	.105	.124	10 & 22	NS	.121
6 & 20	.115	NS	10 & 23	.097	NS
6 & 22	.156	.109	10 & 26	NS	.095
6 & 23	.096	.095	10 & 27	NS	.165
6 & 24	NS	.141	10 & 29	NS	.158
6 & 26	.133	.127	10 & 34	NS	.105
6 & 27	.152	.103			
6 & 31	.103	NS			
	Item No. 14			Item No. 24	
14 & 1	.148	.111	24 & 4	NS	.114
14 & 2	.143	NS	24 & 6	NS	.141
14 & 4	.143	.125	24 & 12	NS	.116
14 & 6	.159	.152	24 & 14	NS	.106
14 & 8	.135	.145	24 & 15	.099	NS
14 & 10	.099	.098	24 & 16	NS	.092
14 & 12	NS	.108	24 & 17	NS	.093
14 & 17	.164	.116	24 & 18	NS	.115
14 & 18	.098	.146	24 & 25	NS	.103
14 & 20	.102	NS	24 & 26	NS	.116
14 & 22	.159	.207	24 & 29	.096	NS
14 & 23	.156	.159	24 & 32	.157	.158
14 & 24	NS	.106	24 & 33	NS	-.091
14 & 26	.121	.157	24 & 34	.157	NS
14 & 27	.142	.151			
14 & 28	NS	.112			

Table 11

Simple and Multiple Correlations Between Predictors and Criteria, from Stepwise Regression Analysis, for Forms A and B of the CCTST

Criterion - Form A	Best Predictor	Simple Correlation	2nd Predictor	Multiple r
Induction	SAT Verbal	.375**	Gender	.391**
Deduction	SAT Verbal	.416**	SAT Quant	.460**
Analysis	SAT Verbal	.277**	(None)	---
Inference	SAT Verbal	.372**	SAT Quant	.416**
Evaluation	SAT Verbal	.438**	SAT Quant	.449**
Total Score	SAT Verbal	.523**	SAT Quant	.547**
Criterion - Form B				
Induction	SAT Verbal	.427**	SAT Quant	.441**
Deduction	SAT Verbal	.509**	SAT Quant	.553**
Analysis	SAT Verbal	.357**	Age	.370**
Inference	SAT Verbal	.507**	SAT Quant	.555**
Evaluation	SAT Verbal	.453**	SAT Quant	.462**
Total Score	SAT Verbal	.594**	SAT Quant	.623**

** $p < .01$